DESIGN OF THE MUC-6 EVALUATION

Ralph Grishman
Dept. of Computer Science
New York University
715 Broadway, 7th Floor
New York, NY 10003, USA
grishman@cs.nyu.edu

Beth Sundheim
Naval Command, Control and Ocean Surveillance Center
Research, Development, Test and Evaluation Division (NRaD)
Code 44208
53140 Gatchell Road
San Diego, California 92152-7420
sundheim@pojke.nosc.mil

Abstract

The sixth in a series of "Message Understanding Conferences", which are designed to promote and evaluate research in information extraction, was held last fall. MUC-6 introduced several innovations over prior MUCs, most notably in the range of different tasks for which evaluations were conducted. We describe the development of the "message understanding" task over the course of the prior MUCs, some of the motivations for the new format, and the steps which led up to the formal evaluation.[1]

THE MUC EVALUATIONS

Last fall we completed the sixth in a series of Message Understanding Conferences, which have been organized by NRAD, the RDT&E division of the Naval Command, Control and Ocean Surveillance Center (formerly NOSC, the Naval Ocean Systems Center) with the support of DARPA, the Defense Advanced Research Projects Agency. This paper looks briefly at the history of these Conferences and then examines the considerations which led to the structure of MUC-6.[2]

The Message Understanding Conferences were initiated by NOSC to assess and to foster research on the automated analysis of military messages containing textual information. Although called "conferences", the distinguishing characteristic of the MUCs are not the conferences themselves, but the evaluations to which participants must submit in order to be permitted to attend the conference. For each MUC, participating groups have been given sample messages and instructions on the type of information to be extracted, and have developed a system to process such messages. Then, shortly before the conference, participants are given a set of test messages to be run through their system (without making any changes to the system); the output of each participant's system is then evaluated against a manually-prepared answer key.

The MUCs have helped to define a program of research and development. DARPA has a number of information science and technology programs which are driven in large part by regular evaluations. The MUCs are notable, however, in that they have substantially shaped the research program in information extraction and brought it to its current state.[3]

[1] Portions of this article are taken from the paper "Message Understanding Conference-6: A Brief History", in *COLING-96, Proc. of the Int'l Conf. on Computational Linguistics.*

[2] The full proceedings of the conference are to be distributed by Morgan Kaufmann Publishers, San Mateo, California; earlier MUC proceedings, for MUC-3, 4, and 5, are also available from Morgan Kaufmann.

[3] There were, however, a number of individual research efforts in information extraction underway before the first MUC, including the work on information formatting of medical narrative by Sager at New York University [3]; the formatting of naval equipment failure reports at the Naval Research Laboratory [1]; and the DBG work by Montgomery for RADC (now

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 1996		**00-00-1996 to 00-00-1996**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Design of the MUC-6 Evaluation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **New York University,Department of Computer Science,715 Broadway, 7th Floor,New York,NY,10003**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996. Sponsored by the Defense Advanced Research Projects Agency.

14. ABSTRACT
The sixth in a series of "Message Understanding Conferences" which are designed to promote and evaluate research in information extraction, was held last fall. MUC-6 introduced several innovations over prior MUCs, most notably in the range of different tasks for which evaluations were conducted. We describe the development of the "message understanding" task over the course of the prior MUCs, some of the motivations for the new format, and the steps which led up to the formal evaluation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

PRIOR MUCS

MUC-1 (1987) was basically exploratory; each group designed its own format for recording the information in the document, and there was no formal evaluation. By MUC-2 (1989), the task had crystalized as one of template filling. One receives a description of a class of events to be identified in the text; for each of these events one must fill a template with information about the event. The template has slots for information about the event, such as the type of event, the agent, the time and place, the effect, etc. For MUC-2, the template had 10 slots. Both MUC-1 and MUC-2 involved sanitized forms of military messages about naval sightings and engagements.

The second MUC also worked out the details of the primary evaluation measures, recall and precision. To present it in simplest terms, suppose the answer key has N_{key} filled slots; and that a system fills $N_{correct}$ slots correctly and $N_{incorrect}$ incorrectly (with some other slots possibly left unfilled). Then

$$recall = \frac{N_{correct}}{N_{key}}$$

$$precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}}$$

For MUC-3 (1991), the task shifted to reports of terrorist events in Central and South America, as reported in articles provided by the Foreign Broadcast Information Service, and the template became somewhat more complex (18 slots). A sample MUC-3 message and template is shown in Figure 1. This same task was used for MUC-4 (1992), with a further small increase in template complexity (24 slots). For MUC-1 through 4, all the text was in upper case.

MUC-5 (1993), which was conducted as part of the Tipster program, represented a substantial further jump in task complexity. Two tasks were involved, international joint ventures and electronic circuit fabrication, in two languages, English and Japanese. In place of a single template, the joint venture task employed 11 object types with a total of 47 slots for the output — double the number of slots defined for MUC-4 — and the task documentation also doubled in size to over 40 pages in length. A sample article and corresponding template for the MUC-5 English joint venture task are shown in Figures 2 and 3. The text shown is all upper case, but (for the first time) the test materials contained mixed-case text as well.

One innovation of MUC-5 was the use of a nested structure of objects. In earlier MUCs, each event had been represented as a single template – in effect, a single record in a data base, with a large number of attributes. This format proved awkward when an event had several participants (e.g., several victims of a terrorist attack) and one wanted to record a set of facts about each participant. This sort of information could be much more easily recorded in the hierarchical structure introduced for MUC-5, in which there was a single object for an event, which pointed to a list of objects, one for each participant in the event.

The sample template in Figure 3 illustrates several of the other features which added to the complexity of the MUC-5 task. The TIE_UP_RELATIONSHIP object points to the ACTIVITY object, which in turn points to the INDUSTRY object, which describes what the joint venture actually did. Within the INDUSTRY object, the PRODUCT/SERVICE slot has to list not just the specific product or service of the joint venture, but also a two-digit code for this product or service, based on the top-level classification of the Standard Industrial Classification. The TIE_UP_RELATIONSHIP also pointed to an OWNERSHIP object, which specified the total capitalization using standard codes for different currencies, and the percentage ownership of the various participants in the joint venture (which may involve some calculation, as in the example shown here). While each individual feature of the template structure adds to the value of the extracted information, the net effect was a substantial investment by each participant in implementing the many details of the task.

MUC-6: INITIAL GOALS

DARPA convened a meeting of Tipster participants and government representatives in December 1993 to define goals and tasks for MUC-6.[4] Among the goals which were identified were

- demonstrating domain-independent component technologies of information extraction which would be immediately useful
- encouraging work to make information extraction systems more portable
- encouraging work on "deeper understanding"

Each of these can been seen in part as a reaction to the trends in the prior MUCs. The MUC-5 tasks, in

Rome Labs) [2].

[4]The representatives of the research community were Jim Cowie, Ralph Grishman (committee chair), Jerry Hobbs, Paul Jacobs, Len Schubert, Carl Weir, and Ralph Weischedel. The government people attending were George Doddington, Donna Harman, Boyan Onyshkevych, John Prange, Bill Schultheis, and Beth Sundheim.

TST1-MUC3-0080

BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) - [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS HE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND INTO A BLUE RENAULT.

HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RELEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT.

LAST WEEK, FEDERICO ESTRADA VELEZ HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

0. MESSAGE ID	TST1-MUC3-0080
1. TEMPLATE ID	1
2. DATE OF INCIDENT	03 APR 90
3. TYPE OF INCIDENT	KIDNAPPING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"THREE HEAVILY ARMED MEN"
6. PERPETRATOR: ID OF ORG(S)	"THE EXTRADITABLES"
7. PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: "THE EXTRADITABLES"
8. PHYSICAL TARGET: ID(S)	*
9. PHYSICAL TARGET: TOTAL NUM	*
10. PHYSICAL TARGET: TYPE(S)	*
11. HUMAN TARGET: ID(S)	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR")
12. HUMAN TARGET: TOTAL NUM	1
13. HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL: "FEDERICO ESTRADA VELEZ"
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPE(S)	*
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17. EFFECT ON PHYSICAL TARGET(S)	*
18. EFFECT ON HUMAN TARGET(S)	-

Figure 1: A sample message and associated filled template from MUC-3 (terrorist domain). Slots which are not applicable to this type of incident (a kidnapping) are marked with an "*". For several of these slots, there are alternative "correct" answers; only one of these answers is shown here.

<DOCNO> 0592 </DOCNO>
<DD> NOVEMBER 24, 1989, FRIDAY </DD>
<SO> Copyright (c) 1989 Jiji Press Ltd.; </SO>
<TXT>
BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE SHIPPED TO JAPAN.
THE JOINT VENTURE, BRIDGESTONE SPORTS TAIWAN CO., CAPITALIZED AT 20 MILLION NEW TAIWAN DOLLARS, WILL START PRODUCTION IN JANUARY 1990 WITH PRODUCTION OF 20,000 IRON AND "METAL WOOD" CLUBS A MONTH. THE MONTHLY OUTPUT WILL BE LATER RAISED TO 50,000 UNITS, BRIDGESTON SPORTS OFFICIALS SAID.
THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID.
BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUB PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES.
WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.
</TXT>
</DOC>

Figure 2: A sample article from the MUC-5 English joint ventures task.

particular, had been quite complex and a great effort had been invested by the government in preparing the training and test data and by the participants in adapting their systems for these tasks. Most participants worked on the tasks for 6 months; a few (the Tipster contractors) had been at work on the tasks for considerably longer. While the performance of some systems was quite impressive (the best got 57% recall, 64% precision overall, with 73% recall and 74% precision on the 4 "core" object types), the question naturally arose as to whether there were many applications for which an investment of one or several developers over half-a-year (or more) could be justified.

Furthermore, while so much effort had been expended, a large portion was specific to the particular tasks. It wasn't clear whether much progress was being made on the underlying technologies which would be needed for better understanding.

To address these goals, the meeting formulated an ambitious menu of tasks for MUC-6, with the idea that individual participants could choose a subset of these tasks. We consider the three goals in the three sections below, and describe the tasks which were developed to address each goal.

SHORT-TERM SUBTASKS

The first goal was to identify, from the component technologies being developed for information extraction, functions which would be of practical use, would be largely domain independent, and could in the near term be performed automatically with high accuracy. To meet this goal the committee developed the "named entity" task, which basically involves identifying the names of all the people, organizations, and geographic locations in a text.

The final task specification, which also involved time, currency, and percentage expressions, used SGML markup to identify the names in a text. Figure 4 shows a sample sentence with named entity annotations. The tag `ENAMEX` ("entity name expression") is used for both people and organization names; the tag `NUMEX` ("numeric expression") is used for currency and percentages.

PORTABILITY

The second goal was to focus on portability in the information extraction task — the ability to rapidly retarget a system to extract information about a different class of events. The committee felt that it was important to demonstrate that useful extraction systems could be created in a few weeks. To meet this goal, we decided that the information extraction task

```
<TEMPLATE-0592-1> :=
    DOC NR: 0592
    DOC DATE: 241189
    DOCUMENT SOURCE: "Jiji Press Ltd."
    CONTENT: <TIE_UP_RELATIONSHIP-0592-1>
<TIE_UP_RELATIONSHIP-0592-1> :=
    TIE-UP STATUS: EXISTING
    ENTITY: <ENTITY-0592-1>
            <ENTITY-0592-2>
            <ENTITY-0592-3>
    JOINT VENTURE CO: <ENTITY-0592-4>
    OWNERSHIP: <OWNERSHIP-0592-1>
    ACTIVITY: <ACTIVITY-0592-1>
<ENTITY-0592-1> :=
    NAME: BRIDGESTONE SPORTS CO
    ALIASES: "BRIDGESTONE SPORTS"
            "BRIDGESTON SPORTS"
    NATIONALITY: Japan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-2> :=
    NAME: UNION PRECISION CASTING CO
    ALIASES: "UNION PRECISION CASTING"
    LOCATION: Taiwan (COUNTRY)
    NATIONALITY: Taiwan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-3> :=
    NAME: TAGA CO
    NATIONALITY: Japan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-4> :=
    NAME: BRIDGESTONE SPORTS  TAIWAN  CO
    LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<INDUSTRY-0592-1> :=
    INDUSTRY-TYPE: PRODUCTION
    PRODUCT/SERVICE: (39 "20,000 IRON AND 'METAL WOOD' [CLUBS]")
<ENTITY_RELATIONSHIP-0592-1> :=
    ENTITY1: <ENTITY-0592-1>
            <ENTITY-0592-2>
            <ENTITY-0592-3>
    ENTITY2: <ENTITY-0592-4>
    REL OF ENTITY2 TO ENTITY1: CHILD
    STATUS: CURRENT
<ACTIVITY-0592-1> :=
    INDUSTRY: <INDUSTRY-0592-1>
    ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>)
    START TIME: <TIME-0592-1>
<TIME-0592-1> :=
    DURING: 0190
<OWNERSHIP-0592-1> :=
    OWNED: <ENTITY-0592-4>
    TOTAL-CAPITALIZATION: 20000000 TWD
    OWNERSHIP-%: (<ENTITY-0592-3> 10)
            (<ENTITY-0592-2> 15)
            (<ENTITY-0592-1> 75)
```

Figure 3: A sample filled template from the MUC-5 English joint ventures task.

```
Mr.  <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
Puris</ENAMEX>, president and chief executive officer of <ENAMEX
TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX
TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX
TYPE="MONEY">$400 million</NUMEX>, but nothing has materialized.
```

Figure 4: Sample named entity annotation.

for MUC-6 would have to involve a relatively simple template, more like MUC-2 than MUC-5; this was dubbed "mini-MUC". In keeping with the hierarchical object structure introduced in MUC-5, it was envisioned that the mini-MUC would have an event-level object pointing to objects representing the participants in the event (people, organizations, products, etc.), mediated perhaps by a "relational" level object.

To further increase portability, a proposal was made to standardize the lowest-level objects (for people, organizations, etc.), since these basic classes are involved in a wide variety of actions. In this way, MUC participants could develop code for these low-level objects once, and then use them with many different types of events. These low-level objects were named "template elements".

As the specification finally developed, the template element for organizations had six slots, for the maximal organization name, any aliases, the type, a descriptive noun phrase, the locale (most specific location), and country. Slots are filled only if information is explicitly given in the text (or, in the case of the country, can be inferred from an explicit locale). The text

> We are striving to have a strong renewed creative partnership with Coca-Cola," Mr. Dooner says. However, odds of that happening are slim since word from Coke headquarters in Atlanta is that...

would yield an organization template element with five of the six slots filled:

```
<ORGANIZATION-9402240133-5> :=
    ORG_NAME: "Coca-Cola"
    ORG_ALIAS: "Coke"
    ORG_TYPE: COMPANY
    ORG_LOCALE: Atlanta CITY
    ORG_COUNTRY: United States
```

(the first line identifies this as organization object 5 from article 9402240133).

Ever on the lookout for additional evaluation measures, the committee decided to make the creation of template elements for all the people and organizations in a text a separate MUC task. Like the named entity task, this was also seen as a potential demonstration of the ability of systems to perform a useful, relatively domain independent task with near-term extraction technology (although it was recognized as being more difficult than named entity, since it required merging information from several places in the text). The old-style MUC information extraction task, based on a description of a particular class of events (a "scenario") was called the "scenario template" task. A sample scenario template is shown in the appendix.

MEASURES OF DEEP UNDERSTANDING

Another concern which was noted about the MUCs is that the systems were tending towards relatively shallow understanding techniques (based primarily on local pattern matching), and that not enough work was being done to build up the mechanisms needed for deeper understanding. Therefore, the committee, with strong encouragement from DARPA, included three MUC tasks which were intended to measure aspects of the internal processing of an information extraction or language understanding system. These three tasks, which were collectively called SemEval ("Semantic Evaluation") were:

- **Coreference**: the system would have to mark coreferential noun phrases (the initial specification envisioned marking set-subset, part-whole, and other relations, in addition to identity relations)
- **Word sense disambiguation**: for each open class word (noun, verb, adjective, adverb) in the text, the system would have to determine its sense using the Wordnet classification (its "synset", in Wordnet terminology)
- **Predicate-argument structure**: the system would have to create a tree interrelating the constituents of the sentence, using some set of grammatical functional relations

The committee recognized that, in selecting such internal measures, it was making some presumptions

regarding the structures and decisions which an analyzer should make in understanding a document. Not everyone would share these presumptions, but participants in the next MUC would be free to enter the information extraction evaluation and skip some or all of these internal evaluations. Language understanding technology might develop in ways very different from those imagined by the committee, and these internal evaluations might turn out to be irrelevant distractions. However, from the current perspective of most of the committee, these seemed fairly basic aspects of understanding, and so an experiment in evaluating them (and encouraging improvement in them) would be worthwhile.

PREPARATION PROCESS

Round 1: Resolution of SemEval

The committee had proposed a very ambitious program of evaluations. We now had to reduce these proposals to detailed specifications. The first step was to do some manual text annotation for the four tasks — named entity and the SemEval triad — which were quite different from what had been tried before. Brief specifications were prepared for each task, and in the spring of 1994 a group of volunteers (mostly veterans of earlier MUCs) annotated a short newspaper article using each set of specifications.

Problems arose with each of the SemEval tasks.

- For coreference, there were problems identifying part-whole and set-subset relations, and distinguishing the two (a proposal to tag more general coreference relations had been dropped earlier); a decision was later made to limit ourselves to identity relations.
- For sense tagging, the annotators found that in some cases Wordnet made very fine distinctions and that making these distinctions consistently in tagging was very difficult.
- For predicate-argument structure, practically every new construct beyond simple clauses and noun phrases raised new issues which had to be collectively resolved.

Beyond these individual problems, it was felt that the menu was simply too ambitious, and that we would do better by concentrating on one element of the SemEval triad for MUC-6; at a meeting held in June 1994, a decision was made to go with coreference. In part, this reflected a feeling that the problems with the coreference specification were the most amenable to solution. It also reflected a conviction that coreference identification had been, and would remain, critical to success in information extraction, and so it was important to encourage advances in coreference. In contrast, most extraction systems did not build full predicate-argument structures, and word-sense disambiguation played a relatively small role in extraction (particularly since extraction systems operated in a narrow domain).

The coreference task, like the named entity task, was annotated using SGML notation. A `COREF` tag has an `ID` attribute which identifies the tagged noun phrase or pronoun. It may also have an attribute of the form `REF`=n, which indicates that this phrase is coreferential with the phrase with ID n. Figure 5 shows an excerpt from an article, annotated for coreference.[5]

Round 2: annotation

The next step was the preparation of a substantial training corpus for the two novel tasks which remained (named entity and coreference). For annotation purposes, we wanted to use texts which could be redistributed to other sites with minimal encumbrances. We therefore selected Wall Street Journal texts from 1987, 1988, and 1989 which had already been distributed as part of the "ACL/DCI" CD-ROM and which were available at nominal cost from the Linguistic Data Consortium.

SRA Corporation kindly provided tools which aided in the annotation process. Again a stalwart group of volunteer annotators was assembled;[6] each was provided with 25 articles from the Wall Street Journal. There was some overlap between the articles assigned, so that we could measure the consistency of annotation between sites. This annotation was done in the winter of 1994-95.

A major role of the annotation process was to identify and resolve problems with the task specifications. For named entities, this was relatively straightforward. For coreference, it proved remarkably difficult to formulate guidelines which were reasonably precise and consistent.[7]

[5] The `TYPE` and `MIN` attributes which appear in the actual annotation have been omitted here for the sake of readability.

[6] The annotation groups were from BBN, Brandeis Univ., the Univ. of Durham, Lockheed-Martin, New Mexico State Univ., NRaD, New York Univ., PRC, the Univ. of Pennsylvania, SAIC (San Diego), SRA, SRI, the Univ. of Sheffield, Southern Methodist Univ., and Unisys.

[7] As experienced computational linguists, we probably should have known better than to think this was an easy task.

```
Maybe <COREF ID="136" REF="134">he</COREF>'ll even leave something from <COREF ID="138"
REF="139"><COREF ID="137" REF="136">his</COREF> office</COREF> for <COREF ID="140"
REF="91">Mr. Dooner</COREF>. Perhaps <COREF ID="144">a framed page from the New York
Times, dated Dec. 8, 1987, showing a year-end chart of the stock market crash earlier
that year</COREF>. <COREF ID="141" REF="137">Mr. James</COREF> says <COREF ID="142"
REF="141">he</COREF> framed <COREF ID="143" REF="144" STATUS="OPT">it</COREF> and kept
<COREF ID="145" REF="144">it</COREF> by <COREF ID="146" REF="142">his</COREF> desk as a
"personal reminder. It can all be gone like that."
```

Figure 5: Sample coreference annotation.

Round 3: dry run

Once the task specifications seemed reasonably stable, NRaD organized a "dry run" – a full-scale rehearsal for MUC-6, but with all results reported anonymously. The dry run took place in April 1995, with a scenario involving labor union contract negotiations, and texts which were again drawn from the 1987-89 Wall Street Journal. Of the sites which were involved in the annotation process, ten participated in the dry run. Results of the dry run were reported at the Tipster Phase II 12-month meeting in May 1995.

An algorithm developed by the MITRE Corporation for MUC-6 was implemented by SAIC and used for scoring the coreference task [4]. The algorithm compares the equivalence classes defined by the coreference links in the manually-generated answer key (the "key") and in the system-generated output (the "response"). The equivalence classes are the models of the identity equivalence coreference relation. Using a simple counting scheme, the algorithm obtains recall and precision scores by determining the minimal perturbations required to align the equivalence classes in the key and response.

THE FORMAL EVALUATION

A call for participation in the MUC-6 formal evaluation was issued in June 1995; the formal evaluation was held in September 1995. The scenario definition was distributed at the beginning of September; the test data was distributed four weeks later, with results due by the end of the week. The scenario involved changes in corporate executive management personnel.

The texts used for the formal evaluation were drawn from the 1993 and 1994 Wall Street Journal, and were provided through the Linguistic Data Consortium. This data had been much less exposed than the earlier Wall Street Journal data, and so was deemed suitable for the evaluation (participants were required to promise not to look at Wall Street Journal data from this period during the evaluation). There had originally been consideration given to using a more varied test corpus, drawn from several news sources. It was decided, however, that multiple sources, with different formats and text mark-up, would be yet another complication for the participants at a time when they were already dealing with multiple tasks.

There were evaluations for four tasks: named entity, coreference, template element, and scenario template. There were 16 participants; 15 participated in the named entity task, 7 in coreference, 11 in template element, and 9 in scenario template. The participants, and the tasks they participated in, are listed in Figure 6.

The results of the MUC-6 evaluations are described in detail in a companion paper in this volume, "Overview of Results of the MUC-6 Evaluation". Overall, the evaluation met many, though not all, of the goals which had been set by the initial planning conference in December of 1993.

The **named entity** task exceeded our expectation in producing systems which could perform a relatively simple task at levels good enough for immediate use. The nearly half the sites had recall and precision over 90%; the highest-scoring system had a recall of 96% and a precision of 97%.

The **template element** task was harder and the scores correspondingly lower than for named entity (ranging across most systems from 65 to 75% in recall, and from 75% to 85% in precision). There seemed general agreement, however, that having prepared code for template elements in advance did make it easier to port a system to a new scenario in a few weeks. The goal for **scenario templates** — mini-MUC — was to demonstrate that effective information extraction systems could be created in a few weeks. Although it is difficult to meaningfully compare results on different scenarios, the scores obtained by most systems after a few weeks (40% to 50% recall, 60% to 70% precision) were comparable to the best scores obtained in prior MUCs.

Pushing improvements in the underlying technology was one of the goals of SemEval and its current

	Task			
site	named entity	coreference	template element	scenario template
BBN Systems and Technology	•		•	•
Univ. of Durham (UK)	•	•	•	•
Knight-Ridder Information	•			
Lockheed-Martin	•		•	•
Univ. of Manitoba	•	•	•	•
Univ. of Massachusetts, Amherst	•	•	•	•
MITRE	•		•	
New Mexico State Univ., Las Cruces	•			
New York Univ.	•	•	•	•
Univ. of Pennsylvania		•		
SAIC	•			
Univ. of Sheffield (UK)	•	•	•	•
SRA	•		•	•
SRI	•	•	•	•
Sterling Software	•		•	
Wayne State Univ.	•			

Figure 6: The participants in MUC-6.

survivor, **coreference**. Much of the energy for the current round, however, went into honing the definition of the task. We may hope that, once the task specification settles down, further evaluations, coupled with the availability of coreference-annotated corpora, will encourage more work in this area.

Appendix: Sample Scenario Template

Shown below is a sample filled template for the MUC-6 scenario template task. The scenario involved changes in corporate executive management personnel. For the text

> McCann has initiated a new so-called global collaborative system, composed of world-wide account directors paired with creative partners. In addition, Peter Kim was hired from WPP Group's J. Walter Thompson last September as vice chairman, chief strategy officer, world-wide.

the following objects were to be generated:

```
<SUCCESSION_EVENT-9402240133-3> :=
    SUCCESSION_ORG:
          <ORGANIZATION-9402240133-1>
    POST: "vice chairman, chief strategy
           officer, world-wide"
    IN_AND_OUT: <IN_AND_OUT-9402240133-5>
    VACANCY_REASON: OTH_UNK
<IN_AND_OUT-9402240133-5> :=
    IO_PERSON: <PERSON-9402240133-5>
    NEW_STATUS: IN
    ON_THE_JOB: YES
    OTHER_ORG: <ORGANIZATION-9402240133-8>
    REL_OTHER_ORG: OUTSIDE_ORG
<ORGANIZATION-9402240133-1> :=
    ORG_NAME: "McCann"
    ORG_TYPE: COMPANY
<ORGANIZATION-9402240133-8> :=
    ORG_NAME: "J. Walter Thompson"
    ORG_TYPE: COMPANY
<PERSON-9402240133-5> :=
    PER_NAME: "Peter Kim"
```

Although we cannot explain all the details of the template here, a few highlights should be noted. For each executive post, one generates a SUCCESSION_EVENT object, which contains references to the ORGANIZATION object for the organization involved, and the IN_AND_OUT object for the activity involving that post (if an article describes a person leaving and a person starting the same job, there will be two IN_AND_OUT objects). The IN_AND_OUT object contains references to the objects for the PERSON and for the ORGANIZATION from which the person came (if he/she is starting a new job). The PERSON and ORGANIZATION objects are the "template element" objects, which are invariant across scenarios.

References

[1] Marsh, E. General Semantic Patterns in Different Sublanguages. In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, R. Grishman and R. Kittredge, eds., Lawrence Erlbaum Assoc., Hillsdale, NJ, 1986.

[2] Montgomery, C. Distinguishing Fact from Opinion and Events from Meta-Events. *Proc. Conf. Applied Natural Language Processing*, 1983.

[3] Sager, N., Friedman, C., and Lyman, M. et al. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, Reading, MA, 1987.

[4] Vilain, M. et al., A Model-Theoretic Coreference Scoring Scheme. *Proc. Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann, San Francisco, 1996.